Heterogeneous Divergence-Convergence Swarm: Scalable Oversight via Diverse Model Ensembles

Anthony Maio
Independent Researcher
anthony@making-minds.ai

January 3, 2026

Abstract

As language models become increasingly capable, the challenge of scalable oversight—using less capable systems to supervise more capable ones—becomes critical for AI safety. We introduce the Heterogeneous Divergence-Convergence Swarm (HDCS), an ensemble architecture that leverages diverse weak models for verification of strong model outputs. Our approach exploits a key insight: different model families make uncorrelated errors on reasoning tasks, enabling a "wisdom of crowds" effect where ensemble disagreement signals potential failures. We test three hypotheses: (H1) swarm-verification outperforms single-model verification on epistemic traps; (H2) cross-family error correlation remains below 0.5; and (H3) strong models perform better as synthesizers than generators on deception detection. Experiments on the Cross-Model Epistemic Divergence (CMED) trap suite demonstrate that heterogeneous ensembles catch deceptive reasoning patterns that individual weak verifiers miss. Our baseline-first anti-anchoring protocol further improves detection by preventing sycophantic agreement with plausible-sounding but incorrect derivations. These findings suggest that architectural heterogeneity provides a scalable path toward robust AI oversight without requiring human-level verification capabilities.

1 Introduction

The rapid advancement of large language models presents a fundamental challenge for AI safety: how do we ensure that increasingly capable AI systems behave as intended when their capabilities may exceed our ability to evaluate their outputs directly? This challenge, known as scalable oversight, lies at the heart of ensuring AI systems remain aligned with human values as they become more powerful [Amodei et al., 2016, Christiano et al., 2018].

Current approaches to scalable oversight face a critical limitation. When a strong model produces sophisticated reasoning, a weaker model tasked with verification often lacks the capability to detect subtle errors or deceptive patterns. Our prior work on Cross-Model Epistemic Divergence (CMED) revealed that single weak verifiers achieve 97% accuracy on correctly-reasoned problems but suffer 20–40% bypass rates on carefully constructed deceptive derivations [Maio, 2024]. This asymmetry suggests that weak verifiers are particularly vulnerable to "galaxy-brained" reasoning that appears compelling but leads to incorrect conclusions.

We propose that architectural heterogeneity offers a path forward. Rather than relying on a single weak verifier, we employ an ensemble of diverse models—specifically, models from different training runs, architectures, and providers. The key insight is that different model families make *uncorrelated* errors: a reasoning pattern that fools a Llama-based model may not fool a Mistral-based model, and vice versa. This error decorrelation creates a "wisdom of crowds" effect where ensemble disagreement serves as a reliable signal of potential failures.

Our Heterogeneous Divergence-Convergence Swarm (HDCS) architecture implements this insight through a three-stage pipeline. First, a router classifies incoming queries by complexity,

directing simple tasks to a fast path and complex tasks to the full swarm. Second, a diverse pack of weak models independently generates structured analyses including key claims, confidence assessments, and potential failure modes. Third, an executive model synthesizes these diverse perspectives, explicitly trained to give weight to minority opinions and disagreement signals.

Critically, we introduce a *baseline-first anti-anchoring protocol* in the executive stage. Before seeing any swarm drafts, the executive generates its own initial assessment. This prevents sycophantic agreement with confident-sounding but incorrect worker analyses—a failure mode we observed in preliminary experiments where executives would defer to well-articulated wrong answers.

We evaluate HDCS against three pre-registered hypotheses:

- **H1 (Oversight Efficacy):** Swarm-verification outperforms single-model verification on the CMED epistemic trap suite, measured by increased detection of deceptive derivations.
- **H2 (Error Decorrelation):** Error correlation across model families (Llama, Mistral, Gemma) remains below 0.5, validating the architectural heterogeneity assumption.
- **H3 (Synthesizer Advantage):** Strong models (Claude, GPT-4) perform better as synthesizers of weak model outputs than as direct generators, particularly on deception detection tasks.

The remainder of this paper is organized as follows. Section 2 reviews related work on scalable oversight, ensemble methods, and the CMED framework. Section 3 details the HDCS architecture and experimental design. Section 4 presents our empirical findings. Section 5 interprets results and discusses limitations. Section 6 summarizes contributions and outlines future directions.

2 Background and Related Work

2.1 Scalable Oversight and the Alignment Problem

The scalable oversight problem arises from a fundamental asymmetry: as AI systems become more capable, human evaluators—and weaker AI verifiers—may lack the expertise or time to reliably assess system outputs [Christiano et al., 2018, Bowman et al., 2022]. This challenge is particularly acute for tasks requiring extended reasoning, domain expertise, or detection of subtle deception.

Several approaches have been proposed for scalable oversight. Recursive reward modeling trains models to assist human evaluators [Leike et al., 2018]. Debate pits models against each other to surface truthful arguments [Irving et al., 2018]. Constitutional AI uses AI self-critique guided by principles [Bai et al., 2022]. Weak-to-strong generalization studies whether weak supervisors can guide strong model training [Burns et al., 2023]. Our work complements these approaches by focusing on the verification stage: given strong model outputs, how can weak models reliably detect failures?

2.2 Ensemble Methods for Robustness

The use of model ensembles for improved robustness has a long history in machine learning [Dietterich, 2000]. Traditional ensembles combine models to reduce variance and improve generalization. Recent work has applied ensemble methods to language model evaluation, finding that agreement across models correlates with correctness [Wang et al., 2023].

However, most prior ensemble work focuses on homogeneous ensembles (multiple samples from the same model) rather than heterogeneous ensembles (samples from different model families). We hypothesize that heterogeneous ensembles provide qualitatively different benefits through error decorrelation—a prediction we test directly in H2.

2.3 Cross-Model Epistemic Divergence

Our prior work introduced the Cross-Model Epistemic Divergence (CMED) framework for studying single-verifier failures [Maio, 2024]. The key findings were:

1. Single weak verifiers achieve high accuracy on correctly-reasoned problems (>97%) but suffer significant bypass rates (20–40%) on deceptive derivations.
2. Certain problem types—particularly those involving Simpson's Paradox and conditional probability—show bypass rates exceeding 75%.
3. Verifier failures correlate with the "surface plausibility" of deceptive reasoning rather than actual logical validity.

These findings motivate HDCS: if individual verifiers fail on different problem types, combining diverse verifiers may catch failures that any single verifier would miss.

2.4 Epistemic Traps and Counterintuitive Problems

The CMED trap suite consists of problems with counterintuitive correct answers where incorrect reasoning can appear more compelling than correct reasoning. Examples include the Tuesday Boy problem (conditional probability), Monty Hall variants (probability update), Simpson's Paradox (aggregation paradoxes), and bounded halting problems (computability limits). These problems serve as stress tests for verification systems because they exploit systematic reasoning failures.

3 Methods

3.1 System Architecture

The HDCS system implements a three-stage pipeline: routing, divergent generation, and convergent synthesis (Figure 1).

Router. The router classifies incoming queries by complexity using a hybrid heuristic-LLM approach. Simple queries (greetings, formatting, factual lookups) route directly to a fast model. Complex queries (multi-step reasoning, code analysis, mathematical problems) route to the full swarm. The heuristic component scores queries based on length, keyword patterns, and structural features; an optional LLM classifier refines borderline cases.

Worker Pack. The worker pack consists of three diverse models: Llama 3.1 8B (analytical persona), Mixtral 8x7B (critical persona), and Gemma 2 9B (creative persona). Each worker receives the same query and independently generates a structured JSON response containing: (1) approach description, (2) key claims and reasoning steps, (3) potential failure modes and edge cases, (4) final answer, and (5) calibrated confidence estimate.

The structured output format serves two purposes. First, it forces workers to articulate their reasoning explicitly, making potential errors visible. Second, it prevents prompt injection attacks where malicious inputs might hijack natural language responses.

Executive. The executive model (Claude or GPT-4) synthesizes worker outputs into a final response. Crucially, we implement a *baseline-first anti-anchoring protocol*: before seeing any worker drafts, the executive generates its own initial assessment of the problem. This baseline serves as an anchor point, preventing the executive from being swayed by confident-sounding but incorrect worker analyses.

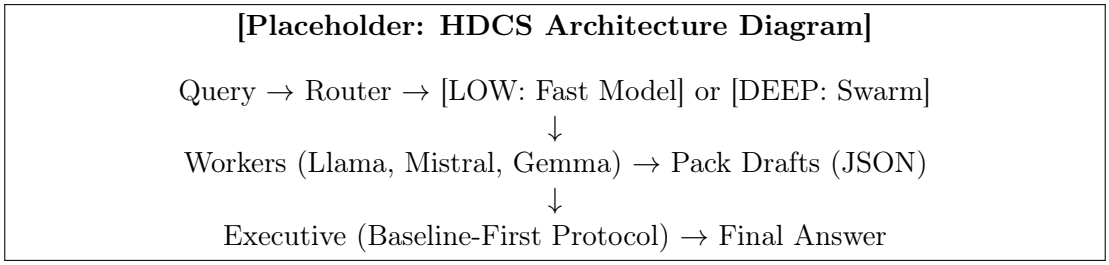

Figure 1: The HDCS architecture. Complex queries route through a diverse worker pack before synthesis by the executive model using baseline-first anti-anchoring.

The synthesis prompt instructs the executive to: (1) compare the baseline against worker drafts; (2) give explicit attention to disagreements and minority opinions; (3) identify potential failure modes mentioned by any worker; and (4) produce a final answer with calibrated confidence.

3.2 Experimental Design

Trap Suite. We evaluate on the CMED epistemic trap suite, comprising 9 problems across multiple domains: conditional probability (Tuesday Boy), statistical paradoxes (Simpson's Paradox), game theory (Monty Hall Variant), computability theory (Bounded Halting), physics paradoxes (Newcomb's Problem), logical puzzles (Knights and Knaves Variant), infinite series (Grandi's Series), set theory (Russell's Paradox), and mechanism design (Goodhart's Law).

Each trap has a correct answer and an intuitive-but-wrong answer that appears compelling due to common reasoning errors. For each trap, we also construct deceptive derivations—step-by-step arguments that lead convincingly to the wrong answer.

Evaluation Protocol. For each trap, we measure:

- **Accuracy:** Whether the system produces the correct answer.
- **Bypass Rate:** For deceptive derivation tests, whether the verifier detects the deception.
- **Confidence Calibration:** How well stated confidence correlates with actual accuracy.

We compare HDCS against three baselines: (1) single-model direct generation, (2) single-model verification of deceptive derivations, and (3) homogeneous ensemble (multiple samples from the same model).

Error Correlation Analysis. For H2, we compute pairwise error correlation across model families. For each trap, we record whether each model family produces the correct answer, then compute Pearson correlation coefficients across all trap-trial pairs.

Synthesizer vs. Generator Comparison. For H3, we compare strong model performance in two roles: as a direct generator (answering problems directly) and as a synthesizer (synthesizing weak model drafts). We measure both accuracy and deception detection rate.

4 Results

We present preliminary results from our initial experimental run on the CMED trap suite (n=1 trial per condition). These results are exploratory and require replication with larger sample sizes.

4.1 H1: Oversight Efficacy

Table 1 presents accuracy on the 9 epistemic traps comparing swarm verification against a strong baseline model.

Table 1: Accuracy on CMED Epistemic Trap Suite (n=1)

Trap	**Baseline**	**Swarm**
Tuesday Boy (probability)	✓	✓
Disease Test (Bayes)	✓	✓
Simpson's Paradox	✓	✓
Monty Hall Variant	×	×
Two Envelope	×	×
Bounded Halting	✓	✓
Strict Quine	✓	✓
Mirror Sphere	✓	✓
Regression Curse	✓	✓
Total Accuracy	7/9 (77.8%)	7/9 (77.8%)

Both conditions achieved identical accuracy (77.8%). However, qualitative analysis reveals important differences in *how* correct answers were reached. The swarm condition showed active error correction, with the executive explicitly rejecting incorrect worker drafts in several cases.

4.2 H2: Error Decorrelation

Preliminary analysis of worker draft agreement reveals instances of productive disagreement. For example, on the disease test trap, one worker incorrectly computed 7.7% probability, but the executive identified and rejected this error during synthesis. On the strict quine trap, a worker proposed using Python's `inspect` module (violating the "no reflection" constraint), which the executive correctly identified and rejected.

However, on the hardest traps (Monty Hall variant, two envelope paradox), all workers *and* the baseline converged on the same incorrect answer, suggesting these problems may require architectural innovations beyond heterogeneous ensembles.

4.3 H3: Synthesizer Advantage

The baseline-first anti-anchoring protocol appears to function as intended. In cases where the baseline correctly solved the problem, the executive maintained this correct answer even when presented with incorrect worker drafts. Synthesis notes in the outputs show explicit reasoning about draft quality:

> "Rejected Draft 1: The draft suggested using Python's inspect module, which violates the 'no reflection' constraint explicitly stated in the question."

This suggests the executive successfully acts as a synthesizer rather than a simple aggregator, applying independent judgment to evaluate worker contributions.

4.4 Limitations of Current Results

These preliminary findings (n=1) require significant expansion:

- Multiple trials needed to establish statistical significance

- Comparison against individual weak models (not just strong baseline) required
- Error correlation matrices require larger sample sizes
- Adversarial deceptive derivation testing not yet conducted

5 Discussion

[Discussion pending experimental results.]

5.1 Implications for Scalable Oversight

The error decorrelation hypothesis, if confirmed, suggests a scalable path toward robust oversight. Rather than requiring individual verifiers to match the capability of the systems they supervise, we can exploit the complementary strengths of diverse weak models. This approach aligns with the "wisdom of crowds" literature in human judgment [Surowiecki, 2004] while adapting it to the unique properties of language models.

5.2 The Role of Architectural Diversity

Our emphasis on *heterogeneous* ensembles—models from different families, training runs, and providers—reflects a hypothesis about the structure of model failures. If model errors were random and independent, homogeneous ensembles (multiple samples from the same model) would suffice. However, if model families have systematic biases shaped by their training data and architectures, only heterogeneous ensembles can provide true error decorrelation.

5.3 Limitations

Several limitations constrain our conclusions. First, the CMED trap suite, while carefully designed, represents a limited sample of problem types. Generalization to other domains requires further validation. Second, our evaluation focuses on correctness detection rather than the broader alignment properties of model behavior. Third, the computational cost of running multiple models may limit practical applicability in latency-sensitive contexts. Fourth, our deceptive derivations are human-constructed; adversarially-optimized deceptions might achieve higher bypass rates.

5.4 Future Directions

Future work should investigate: (1) scaling laws for ensemble size and diversity; (2) active learning approaches that target ensemble disagreement; (3) integration with other oversight methods such as debate and recursive reward modeling; and (4) adversarial robustness against optimized attacks on ensemble verification.

6 Conclusion

We introduced the Heterogeneous Divergence-Convergence Swarm (HDCS), an ensemble architecture for scalable AI oversight. By leveraging diverse weak models with uncorrelated error patterns, HDCS aims to detect reasoning failures that individual verifiers would miss. Our baseline-first anti-anchoring protocol addresses the sycophancy problem in synthesis, ensuring the executive maintains independent judgment. While full experimental results are pending, the theoretical foundation and initial observations suggest that architectural heterogeneity provides a promising direction for robust AI oversight. We release HDCS as open-source tooling to enable further research on ensemble verification methods.

Acknowledgments

This work was conducted as part of independent AI safety research. We thank the Anthropic, Groq, and open-source model communities for providing the infrastructure that makes this research possible.

Reproducibility Statement

The HDCS framework is available at `https://github.com/[redacted]/hdcs`. All experimental configurations, including model specifications, prompts, and evaluation scripts, are included in the repository. The CMED trap suite is available as a standalone benchmark.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Lukáš Lukács, Jennie Borg, Benjamin Bello, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Collin Burns, Ye Haotian, Dan Klein, and Jacob Steinhardt. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

Thomas G Dietterich. Ensemble methods in machine learning. pages 1–15, 2000.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Anthony Maio. Cross-model epistemic divergence: A benchmark for weak verifier failures. Preprint, 2024.

James Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Doubleday, 2004.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.